Scalable Topic Modeling: Online Learning, Diagnostics, and Recommendation
FINAL REPORT

Performing Agency:

David Blei
Columbia University
500 W 120th St
New York, NY
10027

Sponsoring Agency:

Office of Naval Research
875 N Randolph St
Arlington, VA
22217

REPORT

The main activity of my research group is to build and develop the *probabilistic pipeline*. When solving problems with data, we take the following steps.

1. We make assumptions about our data, embedding it in a *probability model* containing hidden and observed random variables.

2. Given observations, we use *inference algorithms* to estimate the conditional distribution of the hidden variables. This is the central statistical and computational problem.

3. With the results of inference, we use our model to form predictions about the future, explore the data, or otherwise apply what we learned to solve a problem.

4. We criticize our model, understand where it went right and wrong, and repeat the process to revise it.

The pipeline cleanly divides the essential activities of data analysis and facilitates collaborative solutions to data science problems. Building models and using them are activities that require domain experts: They tell us what kinds of assumptions they want to make, and how they want to use the results of what we might discover from their data. Inference is a computational and statistics problem. Given the assumptions and data, the problem of estimating the conditional distribution is a well-defined mathematical problem. Model checking and application again requires the domain expert, who can identify what to expect and which areas of the problem are important to success.

For this project, we developed many aspects of this pipeline, particularly around *scalable online learning*, *model checking*, and *recommendation systems*. More broadly, we worked on computational algorithms for fitting models (scalable learning), algorithms for aiding domain experts to build models (model checking),

and real-world applications to test our ideas (recommendation). We went beyond the scope of the proposal in several ways, exploring applications as diverse as neuroscience, sociology, and genetics.

All of our research results are listed at the end of this report. I will highlight several publications of note.

The first is "Stochastic Variational Inference" (JMLR, 2013); this paper scaled up modern Bayesian computation, allowing us to fit many complex models to massive data. In one way, it is the culmination of this project.

The second is "Black Box Variational Inference" (AISTATS, 2014). While stochastic variational inference scaled Bayesian computation up to massive data, black box variational inference expands the scope of scalable Bayesian computation to models that were previously too difficult to work with.

Both of these algorithms, in retrospect, have had a significant impact. They are widely cited and widely *implemented* in open-source software packages. Many of our other publications for this project adapted these ideas including, notably, a paper in *Proceedings of the National Academy of Sciences* (Gopalan et al., 2013) on analyzing massive social networks.

Finally, I point out "Build, Compute Critique, Repeat: Data Analysis with Latent Variable Models" (Annual Review of Statistics, 2014). This is a review article that outlines the full perspective of modern applied probabilistic modeling, including inference, model checking, and applications.

Overall, this project was a success. Between 2011 and 2014, my group has significantly pushed the needle on modern Bayesian machine learning. We have developed new and impactful algorithms, stretched its scope to new applications, and further developed the craft of iterative criticism and model-building.

Refereed Journal Articles

1. D. Blei. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application*, 1 203–232, 2014.

2. S. Gershman, D. Blei, K. Norman, and P. Sederberg. Decomposing spatiotemporal brain patterns into topographic latent sources. *NeuroImage*, 98:91–102, 2014.

3. J. Manning, R. Ranganath, K. Norman, and D. Blei. Topographic factor analysis: A Bayesian model for inferring brain networks from neural data. *PLoS ONE*, 9(5), 2014.

4. P. Gopalan and D. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110 (36) 14534–14539, 2013.

5. M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

6. C. Wang and D. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:1005–1031, 2013.

7. P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41:6, 2013.

8. D. Blei. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1), 2013.

9. D. Blei. Comment on multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108 (503) 771–772, 2013.

10. B. Chen, G. Polatkan, G. Sapiro, D. Blei, D. Dunson, L. Carin. Deep learning with hierarchical convolutional factor analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8), 2013.

11. J. Paisley, C. Wang and D. Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(2):235–272, 2012.

12. D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

13. S. Gershman and D. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12, 2012.

14. D. Blei and P. Frazier. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488, 2011.

15. L. Hannah, D. Blei and W. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

16. S. Gershman, D. Blei, F. Pereira, and K. Norman. A topographic latent source model for fMRI data. *NeuroImage*, 57:89–100, 2011.

Refereed Conference Articles

17. N. Houlsby and D. Blei. A filtering approach to stochastic variational inference. In *Neural Information Processing Systems*, 2014.

18. S. Mandt and D. Blei. Smoothed gradients for stochastic variational inference. In *Neural Information Processing Systems*, 2014.

19. P. Gopalan, L. Charlin, and D. Blei. Content based recommendations with Poisson factorization. In *Neural Information Processing Systems*, 2014.

20. R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

21. P. Gopalan, F. Ruiz, R. Ranganath, and D. Blei. Bayesian nonparametric Poisson factorization for recommendation systems. In *Artificial Intelligence and Statistics*, 2014.

22. M. Rabinovich and D. Blei. The inverse regression topic model. In *International Conference on Machine Learning*, 2014.

23. P. Gopalan, C. Wang and D. Blei. Modeling overlapping communities with node popularities. In *Neural Information Processing Systems*, 2013.

24. D. Kim, P. Gopalan, D. Blei, and E. Sudderth. Efficient online inference for Bayesian nonparametric relational models. In *Neural Information Processing Systems*, 2013.

25. R. Ranganath, C. Wang and D. Blei. An adaptive learning rate for stochastic variational inference. In *International Conference on Machine Learning*, 2013.

26. P. Gopalan, D. Mimno, S. Gerrish, M. Freedman, and D Blei. Scalable inference of overlapping communities. In *Neural Information Processing Systems*, 2012.

27. S. Gerrish and D. Blei. How they vote: Issue-adjusted models of legislative behavior. In *Neural Information Processing Systems*, 2012.

28. C. Wang and D. Blei. Truncation-free online variational inference for Bayesian nonparametric models. In *Neural Information Processing Systems*, 2012.

29. J. Paisley, D. Blei and M. Jordan. Variational Bayesian inference with stochastic search. In *International Conference On Machine Learning*, 2012.

30. D. Mimno, M. Hoffman and D. Blei. Sparse stochastic inference for latent Dirichlet allocation. In *International Conference On Machine Learning*, 2012.

31. S. Gershman, M. Hoffman and D. Blei. Nonparametric variational inference. In *International Conference On Machine Learning*, 2012.

32. A. Chaney and D. Blei. Visualizing topic models. In *International AAAI Conference on Weblogs and Social Media*, 2012.

33. J. Paisley, D. Blei, and M. Jordan. Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, 2012.

34. S. Ghosh, A. Ungureanu, E. Sudderth, and D. Blei. A Spatial distance dependent Chinese restaurant process for image segmentation. In *Neural Information Processing Systems*, 2011.

35. C. Wang and D. Blei. Collaborative topic modeling for recommending scientific articles. In *Knowledge Discovery and Data Mining*, 2011. **Best Student Paper Award.**

36. D. Mimno and D. Blei. Bayesian checking for topic models. In *Empirical Methods in Natural Language Processing*, 2011.

37. S. Gerrish and D. Blei. Predicting legislative roll call from text. In *International Conference on Machine Learning*, 2011. **Distinguished Application Paper Award.**

38. J. Paisley, D. Blei, and L. Carin. Variational inference for stick-breaking beta process priors. In *International Conference on Machine Learning*, 2011.

39. J. Paisley, C. Wang and D. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*, 2011. **Notable Paper Award.**

40. C. Wang, J. Paisley and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2011.